Classification of THz pulse signals using two-dimensional cross-correlation feature extraction and non-linear classifiers

Title: Classification of THz pulse signals using two-dimensional cross-correlation feature extraction and non-linear classifiers

Author: Siuly Xiaoxia Yin Sillas Hadjiloucas Yanchun Zhang

PII: S0169-2607(15)30121-8
DOI: http://dx.doi.org/doi:10.1016/j.cmpb.2016.01.017
Reference: COMM 4070

To appear in: *Computer Methods and Programs in Biomedicine*

Received date: 19-8-2015
Revised date: 20-1-2016
Accepted date: 21-1-2016

Please cite this article as: Siuly, X. Yin, S. Hadjiloucas, Y. Zhang, Classification of THz pulse signals using two-dimensional cross-correlation feature extraction and non-linear classifiers, *Computer Methods and Programs in Biomedicine* (2016), http://dx.doi.org/10.1016/j.cmpb.2016.01.017

Classification of THz pulse signals using two-dimensional cross-correlation feature extraction and non-linear classifiers

Siuly[1], Xiaoxia Yin[1], Sillas Hadjiloucas[2] and Yanchun Zhang[1]

[1]Centre for Applied Informatics, College of Engineering & Science

Victoria University, Melbourne, Australia

[2]School of Systems Engineering, University of Reading, Reading RG6 6AY, UK

siuly.siuly@vu.edu.au; xiaoxia.yin@vu.edu.au; s.hadjiloucas@reading.ac.uk; yanchun.zhang@vu.edu.au

Abstract.

This work provides a performance comparison of four different machine learning classifiers: multinomial logistic regression with ridge estimators (MLR) classifier, k-nearest neighbours (KNN), support vector machine (SVM) and naïve Bayes (NB) as applied to terahertz (THz) transient time domain sequences associated with pixelated images of different powder samples. The six substances considered, although have similar optical properties, their complex insertion loss at the THz part of the spectrum is significantly different because of differences in both their frequency dependent THz extinction coefficient as well as differences in their refractive index and scattering properties. As scattering can be unquantifiable in many spectroscopic experiments, classification solely on differences in complex insertion loss can be inconclusive. The problem is addressed using two-dimensional (2-D) cross-correlations between background and sample interferograms, these ensure good noise suppression of the datasets and provide a range of statistical features that are subsequently used as inputs to the above classifiers. A cross-validation procedure is adopted to assess the performance of the classifiers. Firstly the measurements related to samples that had thicknesses of 2mm were classified, then samples at thicknesses of 4 mm, and after that 3 mm were classified and the success rate and consistency of each classifier was recorded. In addition, mixtures having thicknesses of 2 and 4 mm as well as mixtures of 2, 3 and 4 mm were presented simultaneously to all classifiers. This approach provided further cross-validation of the classification consistency of each algorithm. The results confirm the superiority in classification accuracy and robustness of the MLR (least accuracy 88.24%) and KNN (least accuracy 90.19%) algorithms which consistently outperformed the SVM (least accuracy 74.51%) and NB (least accuracy 56.86%) classifiers for the same number of feature vectors across all studies. The work establishes a general methodology for assessing the performance of other hyperspectral dataset classifiers on the basis of 2-D cross-correlations in far-infrared spectroscopy or other parts of the electromagnetic spectrum. It also advances the wider proliferation of automated THz imaging systems across new application areas e.g., biomedical imaging, industrial processing and quality control where interpretation of hyperspectral images is still under development.

Key-words: Terahertz spectroscopy; 2-D cross-correlation; Multinomial logistic regression classifier; K-nearest neighbours; Support vector machine; Naïve Bayes.

1. Introduction

Over the past 20 years, terahertz (THz or T-rays) pulsed imaging has become an increasingly popular complementary imaging modality due to its ability to simultaneously acquire both spatial and spectral information at a previously inaccessible part of the electromagnetic spectrum [1]. The technique nicely complements existing methods in the XUV, UV, visible and infrared parts of the spectrum. T-rays have a number of unique characteristics, which give rise to a large number of potential applications in very diverse fields such as, security, pharmaceutical quality control, medical imaging and material science [2]. In addition, owing to their low photon energy, T-rays are non-ionizing and are thus considered of not inducing damage to tissue or DNA. Therefore, they are currently considered as viable alternatives to X-rays for imaging in biomedical applications where the subject may not be irradiated by X-rays e.g., for mammograms in pregnant or lactating women. Alternative applications benefitting from this technology include retection (detection of hidden objects or substances within a package), where THz tomographic image contrast can be superior to conventional methods such as X-rays that only differentiate objects or regions in an image mainly on the basis of different sample density but have difficulties in detecting plastic objects or soft biological materials of similar density. In contrast, T-ray wavelengths can pass through dry substances (e.g. thin cardboard and plastics), as well as through non-polar, non-metallic materials and can show spectral differences due to a different extinction coefficient between samples. Concealed weapons or products contained in plastic packages and non-metallic components that are not readily detectable by other means can therefore be easily detected using THz imaging techniques. The approach is also particularly promising for the detection of specific chemical and biological agents [3, 4], through chemical fingerprinting. Within a pharmaceutical setting, such systems can perform multiple functions [5] enabling the identification of drug polymorphisms [6], providing information on coating structures [7-10], enabling the identification of phase transitions in chemical compounds [11] or degree of substance crystallinity [12-15] providing opportunities for tailoring the formulations at each processing step or enabling the monitoring of physicochemical product deterioration during processing or storage [16-17]. Furthermore, the high transparency of polymer materials to THz waves enables non-destructive inspection of encapsulated substances such as drugs [18], making this imaging modality particularly useful to the pharmaceutical industry. It is therefore clear that quality control for pharmaceutical industry is therefore seen as a potentially important application area for THz imaging systems [19-22] provided reliable machine learning techniques can be integrated with the sensing equipment. The use of T-ray pulse transients for simultaneously extracting information on densities, thicknesses and

number of absorber molecules per unit volume in different powder samples forms the basis for simultaneously addressing detection and classification requirements across both pharmaceutical [23-25] as well as security industries [25-26].

It is worth noting that THz imaging spectrometers excite samples with femtosecond duration pulses which are extremely broadband, where a pulse spectrum spans over a frequency range between 100 GHz (such excitation is associated with a wavelength of 3 mm) up to 3 THz (with a corresponding wavelength of 0.1 mm) and in some systems all the way up to 10 THz (with a corresponding wavelength of 0.03 mm). As a consequence, many experiments may also contain spectral signatures associated with measurement artefacts at the Rayleigh to Mie transition region where the excitation wavelength becomes similar to the size of the particles that need to be characterized. As a consequence, in all femtosecond pulse based THz imaging systems it is not uncommon that measurements of many powdered samples can miss out a scattering component of the THz radiation, especially at frequencies closer to the infrared part of the spectrum. Scattering can cause particularly severe problems in THz time domain spectrometry, such instruments are only reliable at measuring transmittance (by measuring attenuation), or reflection (impedance mismatch) within a well-defined aperture, at a well-defined sample-air interface and across a single plane defined perpendicularly to the direction of propagation of the THz pulse. From these measurements absorption can finally be estimated, under the provision that scattering is negligible. In the datasets chosen to be investigated in the current study, there is some unquantifiable by other means scattering component because the samples have grains of different dimensions, hence there is a problem in adopting standard processing and perform classification solely based on information associated with specific spectral features. Since THz pulse imaging is extremely broadband, there may be different degree of scattering associated with the spectral signatures across different spectral bands, this is especially true if samples are in powdered form. Such problems may further be exacerbated if the powdered sample is elliptical in shape [27]. As a result one would expect different degree of deviation of the obtained absorption results associated with complex insertion loss measurements at different spectral bands and the calculated spectral extinction coefficient may significantly deviate from its true value. Finally, contrary to continuous wave based measurement systems [28], pulse transient systems spatially focus the THz radiation dramatically so as to improve on the signal-to-noise ratio during the measurement process, this has additional adverse effects in that there are deviations in the extraction of the complex insertion loss function which requires an assumption that an angular spectrum of plane waves is incident on the sample, clearly such

focusing can lead to additional systematic errors in estimating the complex insertion loss function while also exacerbates the effects of scattering as sample excitation takes place over a range of angles across the sample aperture; such angular dependency of the degree of scattering makes also collection of scattered energy difficult to perform and quantify [29]. These problems lead to a need for reassessment of what can be considered as useful features that can be meaningfully extracted in a THz imaging experiment so that an automated machine learning methodology for the classification of samples using THz imaging systems can be developed.

A further aim of the proposed approach is to preserve compatibility with other de-noising techniques. Typically, the THz pulse signals contain noise due to both systematic and random errors and thus the signal-to-noise ratios in the acquired THz spectra are low. This introduces significant problems in the analysis and interpretation of spectra as well as the classification of samples (there are collinearity issues at spectral bands where the signal to noise ratio is low, such collinearity results in spikes in the error because calculation of the complex insertion loss is based on a ratiometric process). It is therefore often the case that the acquired complex insertion loss signatures may contain limited discriminative information. One method to reduce errors due to noise is to co-average subsequent measurements for the same pixel, however this dramatically increases the time required to perform the measurement, with several images reported in the current literature being acquired over a period of several minutes or even several hours. Such approach also does not address spectral bands where the source output spectral power is low.

Although there is an extensive literature on the signal processing of THz spectra, 2-D cross-correlation techniques [30-32] have attracted less attention despite their de-noising or feature extraction potential. Such approach represents a natural extension of existing THz deconvolution approaches [33] and complements de-noising algorithms using auto-regression with exogenous inputs (ARX) and subspace approaches [34-35], or other state-of-the-art signal analysis approaches [e.g.36, 37]. It is also interesting to note that cross-correlations are extensively used in different spectral bands (XUV, UV, visible, infra-red) but are not as widespread within the THz community. By performing a 2-D cross correlation between the sample and background time domain signals, excellent de-noising is achieved while preserving any phase differences (which are associated with the dispersion of the sample) that might be present between the two signals. The obtained cross-correlogram is a nearly noise-free signal that can convey superior discriminative phase information compared to the original time domain interferogram signal [30].

In recent years, a number of methods have been proposed for feature extraction in conjunction with sample classification on the basis of THz pulsed signatures. Most recently, Yin *et al.*, [38] used directly both the real as well as complex values associated with the Fourier Transform (FT) of the corresponding time domain signatures to perform de-noising and sample classification. Furthermore, in [39], Yin *et al.*, established that it is possible to use specific features form the Fourier spectrum of the sample to extract T-ray feature sets for binary and multi-class classification. The general approach in that method is based on selecting specific feature vectors in the frequency-domain by taking the FT after de-convolving the measured signals with a reference pulse. Alternative feature extraction algorithms using adaptive wavelet coefficients in conjunction with ARX, ARMAX as well as subspace algorithms for signal de-embedding have also been suggested in the THz literature, confirming the merits of this approach [34, 35, 40]. These measurements, however, were not performed on powdered samples but on samples having uniform thickness or well controlled thickness (micro-spectroscopy using waveguides). Furthermore, in order to use information associated with the dispersion of the sample in conjunction with the molecular extinction coefficient and number of absorbers across the spectrum of the measurements, alternative classification approaches making use of the discrete wavelet transforms (DWT) in T-ray measured powder samples have also been reported [41]. The goal to further reduce the input vector of the classifier so as not to compromise its generalization ability has led to the development of a hybrid pre-processing algorithm that used Auto Regressive (AR) modelling within the wavelet decomposed sub-bands of the THz pulsed signals [24]. The work complemented previous attempts by Ferguson *et al.*, [37] to classify powders concealed within envelopes, despite the presence of strong scattering. To our knowledge these studies and the extreme learning approach recently developed [38] are the only ones that combine advanced signal pre-processing with classification for powdered samples imaged using THz transient spectrometry. In addition, to the best of our knowledge, 2-D cross-correlation techniques have never been used for feature extraction of T-ray spectra of powdered samples. Finally, within an Analytical Chemistry context, cross-correlation techniques are not usually explored within a machine learning perspective but are mainly discussed as a viable de-noising tool or to elucidate fast transient processes observed using pump-probe techniques. This differentiates the current study as it is focused in advancing current algorithms from a machine learning perspective. Such considerations have led us to develop the proposed methodology.

A further aim of the work is also to assess the potential of combining 2-D cross-correlation at the pre-processing feature extraction step while systematically assessing its impact to the performance of different classifiers. This is achieved by focusing the investigations on the identification of several powder samples of different composition. Our goal is to demonstrate a generic feature extraction approach that fully utilizes the different characteristic features found in THz pulse signals so that may be used with minimal reformulation across different T-ray data sets. Such approach paves the way towards the development of a suitable machine learning classification algorithm that could be reliably used to identify different materials independent of their thickness on the basis of their estimated spectrally dependent extinction coefficient even in the presence of some unquantifiable scattering. Feature extraction is the most crucial step in this type of pattern recognition because the classification performance will be significantly degraded if the features are not chosen wisely [42]. A further aim is to reduce the extracted features to prevent over fitting while retaining most of the useful information residing in the original vector. In order to reduce the dimensionality of the cross-correlation sequences, it is also proposed that ten statistical features are extracted from each cross-correlation sequence. The validity of the cross-correlogram features as preferred inputs is subsequently evaluated in a systematic manner by considering four machine learning algorithms: multinomial logistic regression classifier with ridge estimators (MLR), k-nearest neighbours (KNN), support vector machine (SVM) and naïve Bayes (NB). The choice of these classifiers is based on their simplicity and effectiveness in their implementation. Investigations are performed to test both multi-class as well as binary classification of T-ray pulse transmission signals. A 10-fold cross-validation method is used for assessing the performance of the proposed methodology. This procedure divides the feature vector sets into ten approximately equal-sized distinct partitions. One partition is used for testing, whereas the other partitions are used for training the classifiers. To further improve the estimate, the procedure was repeated 10 times and all performance metrics over these runs are averaged. The average performances associated with the test data is then adopted as the preferred overall performance evaluation criterion. The investigations aim to elucidate which one of the four classifiers would consistently achieve the most reliable classification. The powder samples used in the study have similar optical properties but different composition and different complex insertion loss at the THz part of the spectrum.

The paper is organized as follows: Section 2 provides an overview of the algorithm adopted to perform the cross-correlation process, details of the statistical feature extraction

process, feature aggregation and cross-validation as well as a brief outline of the methods associated with the four classifiers. This section also provides information regarding the nature of the datasets. In Section 3, the application of the 2D cross-correlation procedure to the THz datasets is discussed. The selection of optimum parameter values for the reported classifiers and the performance evaluation criteria are also discussed in this section. A performance comparison of all the classifiers is presented and discussed in Section 4. Finally Section 5 draws some conclusions and provides directions for further research.

2. Proposed classification methodology

2. 1. Overview of the pre-processing and classifier design

The general classifier structure consists of four main processing blocks: computation of cross- correlation sequence, statistical feature extraction, feature aggregation and cross validation and classifier decision observation. The 2-D cross correlation technique extracts the information from the T-ray pulsed signals and acquires cross-correlation sequences from each sample class. In this study, each powder substance is considered to belong to a single class: sand (class 1), talcum (class 2), salt (class 3), powdered sugar (class 4), wheat flour (class 5), and baking soda (class 6). The sample holder (free-space equivalent of a cuvette) signal is the reference signal used for evaluating the complex insertion loss. Using the reference signal in conjunction with the other sample signal in a class, a cross-correlation sequence is computed on a pixel by pixel basis by the 2-D cross correlation. Once the characteristic features are extracted from each cross-correlation sequence associated with every class, all features are integrated forming a feature set. Following this process, cross-validation is applied to generate training and testing sets for evaluation. The detection stage identifies the several powder categories on the basis of the feature sets. Finally classifier decisions are observed.

2.1.1. Computation of cross-correlation sequences

The 2-D cross-correlation technique [30,43-44] is used to calculate a cross-correlation sequence (denoted by '*CC(k,l)*') between the reference signal and any other signal belonging to a distinct class. The graphical presentation of a cross-correlation sequence is commonly known as a cross-correlogram. The 2-D cross-correlation of X (M-by-N matrix) and H (P-by-Q matrix) is a matrix CC of size $(M+P-1) \times (N+Q-1)$:

$$CC(k,l) = \sum_{m=0}^{M-1}\sum_{n=0}^{N-1} X(m,n)\bar{H}(m-k,n-l); \quad -(P-1) \le k \le M-1; -(Q-1) \le l \le N-1 \qquad (1)$$

where X is considered as the reference signal and H is regarded as any other signal belonging to a class of T-ray pulsed signals. The bar over H denotes complex conjugation. The output matrix, $CC(k,l)$, has negative and positive row and column indices. A negative row index corresponds to an upward shift of the rows of H. A negative column index corresponds to a leftward shift of the columns of H. A positive row index corresponds to a downward shift of the rows of H. A positive column index corresponds to a rightward shift of the columns. It is worth mentioning that if each of the signals, X and H, consist of a finite number samples S, the resultant cross-correlation sequence has $2S$-1 samples.

The THz transient transmission reference signal is considered as noiseless for most parts of the spectrum, so the variance in the noise when ratioing a sample with a background does not get disproportionally amplified [45]. Each powder sample is considered as belonging to a distinct class. Fig. 1 illustrates how a cross-correlogram is obtained from a reference signal (holder) and any of the other sample signals, on the basis of Eq. (1).

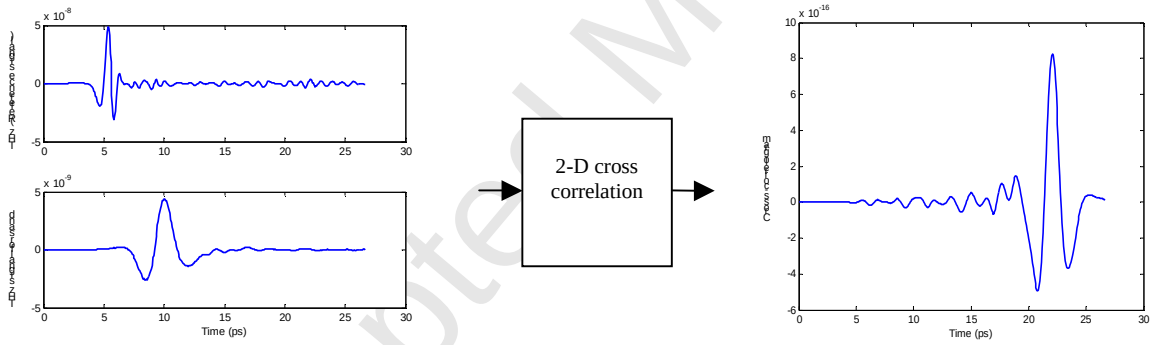


Fig. 1. Typical cross-correlogram from THz background and sample time domain signatures

The cross-correlogram signals convey greater information than the original powder spectra of the sample and reference signals and thus have superior signal to noise ratio than the original signals. In addition, cross-correlograms contain additional information regarding the spectral coherence of the waveforms. As the cross-correlation sequences contain a large number of data points, these needs to be further compressed into a more parsimonious feature space so as not to overwhelm the classifier.

2.1.2. Statistical feature extraction

In order to reduce the dimensions of the cross-correlation sequences, this study considers ten statistical features. These are: mean, standard deviation, skewness, kurtosis, 1st quartile *(Q_1)*,

3rd quartile *(Q_3)*, inter-quartile range *(IQR)*, median, maximum and minimum that are calculated from each cross-correlation sequence. This information is used to create the feature vector sets. There are several valid reasons for the considerations of these ten quantitative feature descriptors. Mean and standard deviation are particularly informative in describing a distribution [46-47]. Skewness provides information on the degree of asymmetry of the observed distribution around its mean [43]. Kurtosis provides a measure of flatness relative to a normal distribution. Q_1 and Q_3, measure how the data are distributed in the two sides of the median. *IQR* is the difference between Q_3 and Q_1 that is used in measuring the spread of a data set, such information can be used to exclude outliers [48-49]. Median which is associated with the observation encountered most often is also an additional valuable metric that needs to be retained for classification purposes. Maximum and minimum values are also used to describe the range of observations within the distribution. Each of the above subroutines is run for each cross-correlation sequence associated with each powder substance. All ten statistical features from each cross correlation sequence and each powder substance form the content of a feature set that is finally associated with each powder material.

2.1.3. Feature aggregation and cross validation

In this stage, the obtained feature set from each powder material are combined to form a composite feature set that contains all the features from all T-ray pulse signals of each powder substance. This feature set is used to generate training and testing sets through the cross- validation process. In order to reduce any bias of training and test data, a *k*-fold cross-validation technique is employed [48, 50, 51] setting k=10. This technique is implemented to create the training set and testing set for evaluation. Generally, with *k*-fold cross validation, the feature vector set is divided into *k* subsets of (approximately) equal size. The proposed classifiers are trained and tested *k* times. Each time, one of the subsets from training is left out. One of the subsets (folds) is used as a test set and the other *k*-1 subsets (folds) are put together to form a training set. Then the average accuracy across all *k* trials is computed to assess the performance of the classifier.

2.1.4. Overview of THz pulse signal classifier algorithms

In the following section, the utility of the calculated feature sets is evaluated through four well established machine learning classifiers: multinomial logistic regression classifier with ridge

estimators (MLR), k-nearest neighbours (KNN), support vector machine (SVM) and naïve Bayes (NB). Overviews of the adopted algorithms are provided below.

MULTINOMIAL LOGISTIC REGRESSION CLASSIFIER WITH RIDGE ESTIMATORS (MLR)

Ridge estimators are used in multinomial logistic regression to improve the parameter estimates and to diminish the error made by further prediction when the application of maximum likelihood estimators (MLE) is inappropriate because of the non-uniqueness of the solution in the data fitting process. When the number of explanatory variables are relatively large and / or when the explanatory variables are highly correlated, the estimates of parameters are unstable, and are not uniquely defined (some are infinite) so the maximum of log-likelihood is achieved at 0 value [52, 53]. In this situation, ridge estimators are used to generate finiteness and uniqueness of MLE to overcome such problems. The above rationale provides the necessary justification for considering the use of such classifier to the current task. For a response variable $Y \in \{1,2,...,k\}$ with k possible values (categories), there are k classes for n instances with m attributes (explanatory variables), the parameter matrix B that requires to be calculated will have dimension $m\times(k\text{-}1)$. In this case, the probability for class j with the exception of the last class is given from:

$$P_j\left(X_i\right)=\frac{exp\left(X_iB_j\right)}{\sum_{j=1}^{k} exp\left(X_iB_j\right)+1} \tag{4}$$

The last class has a probability of occurring given by:

$$1-\sum_{j=1}^{k-1} P_j\left(X_i\right)=\frac{1}{\sum_{J=1}^{K-1} exp\left(X_iB_j\right)+1} \tag{5}$$

and the (negative) multinomial log-likelihood is given from:

$$\mathrm{L}= L=-\sum_{i=1}^{n}\sum_{j=1}^{k-1}\left(Y_{ij}\times In\left(P_j\left(X_i\right)\right)\right)+\left(1-\sum_{j=1}^{k-1} Y_{ij}\right)\times In\left(1-\sum_{j=1}^{k-1} P_j\left(X_i\right)\right)+ridge\times B^2 \tag{6}$$

In order to find the matrix B for which L is minimised, a Quasi-Newton method is used to search for the optimized values of the $m\times(k\text{-}1)$ variables [52]. At this stage it is worth noting that in the current implementation of the algorithm, before we use the optimization procedure, we 'squeeze' the matrix B into a m× (k-1) matrix. A more detail description of the MLR adopted can be found in [52, 53]. In the current study, X indicates the obtained feature

set associated with the six powder substances and *Y* denotes the different categories associated with the six the powder substances.

K-NEAREST NEIGHBOURS (KNN) CLASSIFIER

The rationale for choosing the use of a KNN algorithm is based on the fact that it is a very intuitive method in which the classifier labels the observations based on their similarity in the training dataset. Among the various methods of supervised statistical pattern recognition, the KNN rule is known to achieve consistently high performance, without *a priori* assumptions regarding the distributions from which the training examples are drawn [54]. Given a query vector x_0 and a set of N labelled instances $\{x_i, y_i\}_1^N$, the task of the classifier is to predict the class label of x_0 on the predefined P classes. The KNN classification algorithm tries to find the k nearest neighbours of x_0 and uses a majority vote to determine the class label of x_0. Without prior knowledge, the *KNN* classifier usually evaluates Euclidean distances as a metric [55]. An appropriate value should be selected for k, because the success of classification is very much dependent on this value. There are several methods to choose the k-value; a well-established practical approach is to run the algorithm many times with different k-values (k =1, 2,…, 20), and choose the one with the best performance. A detailed discussion of this method can be found in [56-57]. In the current investigation, we consider the feature vector associated with the powder sample datasets as $\{x_i\}$ and the six powder categories as class label $\{y_i\}$.

SUPPORT VECTOR MACHINE (SVM) CLASSIFIER

The SVM is most popular machines learning tool that can classify data separated by non-linear and linear boundaries, originated from Vapnik's statistical learning theory [58]. The main concept in all SVM algorithms is to first transform the input data into a higher dimensional space and then construct an optimal separating hyper-plane (OSH) between the two classes in the transformed space [39,59]. Those data vectors nearest to the constructed line in the transformed space are referred to as the support vectors. SVM algorithms belong to the more general area of "structural risk minimization" algorithms which have been developed specifically to attain a low probability of generalization error. Because of their versatility and universal applicability to a variety of classification tasks, they have also been considered in the current study. In order to solve nonlinear problems, when the data are not

linearly separable, SVMs usually adopt a nonlinear kernel function [39, 59], which allows better fitting of the hyperplane to the datasets that need to be classified. Recently, SVMs have also been extended to solve multi-class classification problems. One frequently used method in practice is to use a set of pair-wise classifiers, based on one-against-one decomposition [39]. The decision function for binary classification is given from:

$$f(x) = \mathrm{sgn} \sum_{i=1}^{s} y_i \alpha_i k(x_i, x) + b \quad ; 0 < \alpha_i < C \tag{7}$$

where, sgn is the signum function, $K(x_i, x)$ is a kernel function and b is the bias of the training samples. In this work, a radial basis function (RBF) kernel is considered as a choice for identifying different categories of T-ray signals because this was found to give the best classification performance. Here C is the regularization parameter used to tune the trade-off between minimizing empirical risk (e.g. training error). In the current work, the complexity of the machine $C = \frac{N}{\sum_{i=}^{N} K(x_i, x)}$ is always set to its default value, where N denotes the size of the training set, x_i indicates the i^{th} input feature vector set (with a dimensionality of 6) and y_i (i=1,2,..6) is the class label of x_i, containing one of six categories of powder substances.

In the multiclass problem, SVM classification is performed using a collection of decision functions f_{kl}. Here kl indicates each pair of classes selected from separated target classes. The class decision can be achieved by summing up the pairwise decision functions [39].

$$f_k(x) = \sum_{i=1}^{n} \mathrm{sgn}\left(f_{kl}(x)\right) \tag{8}$$

Here n refers to the number of separated target classes. The algorithm proceeds as follows: first assign a label to the class: arg max $f_k(x)$, (k=1,2,…,n). In the above equation, the signum function (sgn) is used to denote a hard threshold decisions [39] i.e., $\mathrm{sgn}\left(f_{kl}(x)\right) = \begin{matrix} 1 & f_{kl}(x) > 0 \\ -1 & f_{kl}(x) \le 0 \end{matrix}$.

The pairwise classification then converts the n-class classification problem into $n(n-1)/2$ two-class problems which cover all pairs of classes. An overview of SVM pattern recognition techniques associated with the proposed methodology may be found in [39, 58, 59].

NAIVE BAYESIAN (NB) CLASSIFIER

The NB is chosen for the current study as it is a straightforward and frequently used probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions [60-62].

The NB classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. Depending on the precise nature of the adopted probability model, the NB classifier can be trained very efficiently in a supervised learning setting. In practical applications, parameter estimation for naive Bayes models uses the method of maximum likelihood. In this classifier, each class with highest post-probability is addressed as the resulting class.

Suppose, $X=\{X_1, X_2, X_3,.....,X_n\}$ is a feature vector set that contains C_k (k=1,2,..m) classes of data to be classified. Each class has a probability $P(C_k)$ that represents the prior probability of identifying a feature into C_k and the values of $P(C_k)$ can be estimated from the training dataset. For the n feature values of X, the goal of classification is clearly to find the conditional probability $P(C_k|\, x_1, x_2, x_3,.....,x_n)$. By Bayes's rule, this probability is equivalent to

$$P(C_k|\, X_1, X_2, X_3,.....,X_n)=\frac{P(C_k)P(\mathrm{X_1}, \mathrm{X_2}, \mathrm{X_3},.....,\mathrm{X_n}|\mathrm{C_k})}{P(C_k)P(\mathrm{X_1}, \mathrm{X_2}, \mathrm{X_3},.....,\mathrm{X_n}|\mathrm{C_k})} \tag{9}$$

The final decision rule for the NB classifier is:

$$classify\,(X_1, X_2, ..., X_n) = \arg\max_{C_k} p(C_k)\prod_{i=1}^{n} P(X_i \,|\, C_k) \tag{10}$$

In the current study, we used the obtained feature vector set as the input in equation (10) and C_k (k=1,2,..6) indicates the number of the six powder categories that the data had to be classified. In the training stage, $P(X_i|C_k)$ is estimated with respect to the training data. In the testing stage, based on the posterior probability $P(C_k|X_i)$, a decision whether a test sample belongs to a class C_k is made. A detailed description of the method can be found elsewhere [54, 60-62].

2.2. Details of the THz sample datasets

The current study explores the ability of T-ray spectroscopy to detect different densities, thicknesses, and concentrations of specific powder samples. This is a powder recognition task for six different powdered substances of 2 mm and 4mm thickness where their spectroscopic signature needs to be de-convolved from that of the holder. The powders are: sand, talcum, salt, powdered sugar, wheat flour, and baking soda. In addition, we also explore the classification fidelity attained for a mixture of 2mm and 4mm thickness samples across all powder substances.

In order to further assess the performance and consistency of the proposed methods, data from 3mm thickness powder samples for the same six powder substances is also considered in this study. The 3mm thickness powder samples have the same composition as

their corresponding 2mm and 4mm thickness powder sample datasets. A well set-up T-ray imaging system which generates femtosecond duration terahertz pulses is used to detect the T-ray sample responses [36, 39]. The 2-D T-ray image of the sample is obtained after separately recording the sample holder transmittance and then inserting the powder sample. The geometry of the experiment preserves the ambiguities associated with the effects of different scattering paths and minor variations in powder thickness across the aperture (pseudo-coherence effects) and density due to slightly different compaction levels across the six substances observed. Sample transmittance is recorded by broadband time-domain THz transient spectrometry. The reported measurements have been conducted at the University of Adelaide Australia [39]. A detailed description of the dataset acquisition process using the THz imaging spectrometer can be found in [38-39, 37].

3. Systematic evaluation of the classifier performance

To systematically evaluate the performance of the proposed 2-D cross-correlation based machine learning algorithms, THz time-domain spectra from all six known powder substances were used. These samples had very similar optical properties but different absorption features at the THz part of the spectrum. The classification task was to correctly identify the specific powders given they had unknown density, thickness and concentration. A preliminary exploration of different powder recognition tasks was first conducted with 2 mm and 4 mm thickness samples. Collected spectra incorporated the distortion from the sample holder, this signature was eliminated by assigning the holder spectrum in the experiments as background (reference) and ratioing the powdered sample spectrum with that of the background so as to extract the complex insertion loss. The following investigations were carried out: (i) multiclass classification of the six categories of powder samples at a thickness of 2mm; (ii) multiclass classification of the six categories of powder samples at a thickness of 4mm (iii) binary classification in each powder substance for a mixture of 2mm and 4mm thickness samples. In order to obtain a further assessment of the consistency of the proposed methodology, we performed the multiclass classification of the six categories of powder samples at a thickness of 3 mm and also evaluated the success of the algorithm to perform multiclass classification in each powder substance for a mixture of 2mm, 3mm and 4mm thickness samples. All the powder sample classification runs were performed using the MATLAB version R2013b software on a personal computer running Windows 7 with an Intel(R) Core(TM) i5-4570S CPU (2.90 GHz) and 8 GB of memory. The following four classification algorithms were used: MLR, KNN, SVM and NB implemented in WEKA

machine learning toolkit [63]. LIBSVM (version 3.2) [64] is used for the SVM classification in WEKA.

3.1. Selection of optimal parameter values for the adopted classifiers

In the MLR method, the parameters are obtained automatically through the ridge estimator. The KNN model has only one parameter *k* which refers to the number of nearest neighbors. By varying *k*, the model can be made more flexible. In the current study, we have chosen the appropriate *k* value through an automatic process following a *k* selection error log as there is no simple rule for selecting *k*. We consider the range of *k* values between 1 and 20, and picked an appropriate *k* value that results in lowest error rate as this is associated with the best model. In the experimental results, we obtain the lowest error rate for *k*=1. For the SVM, the RBF kernel function was employed as an optimal kernel function over several different kernel functions that were tested. As there are no specific guidelines to set the values of the parameters for the MLR and the SVM classifiers, we considered the parameter values that have been used in WEKA as default parameter settings. The NB consists of number of parameters that are estimated from the training examples. Parameter estimation for the NB models uses the method of maximum likelihood.

3.2. Performance evaluation criteria

In this study, we assess the performance of the proposed classifiers using widely accepted metrics such as accuracy, true positive rate (TPR) (also called sensitivity or recall), false positive rate (FPR) (also called false alarm rate or (1-specifity)), precision (also called positive predictive value), F-measure, mean absolute error (MAE) and kappa statistics. These criteria were applied to assess all extracted feature data. The evaluation metric adopted is accuracy rate as percentage of correct prediction [65-67]. The TPR provides the fraction of positive cases that are classified as positive [49, 68]. The FPR [49, 69] is the percentage of false positives predicted as positive from samples belonging to the negative class. The FPR usually refers to the expectancy of the false positive ratio. Precision is a measure which is used to estimate the probability that a positive prediction is correct. F-measure is a combined measure for precision and recall calculated as 2*Precision*Recall /(Precision + Recall) [49]. Mean absolute error (MAE) is used to measure how close predictions are to the eventual outcomes [49]. Kappa is a chance-corrected measure of agreement between the classifications and the true classes [49, 70]. It's calculated by taking the agreement expected by chance away from the observed agreement and dividing by the maximum possible agreement.

3.3. Evaluation of the 2-D cross correlation pre-processing step

The images of powder samples consist of 6 × 51 = 306 pixels. For each pixel, the number of samples associated to a pulse time transient is set to 401. Fig. 2 (a) and (b) shows the time domain responses associated with the THz transmittance of the powdered samples with 2 mm and 4 mm thickness,

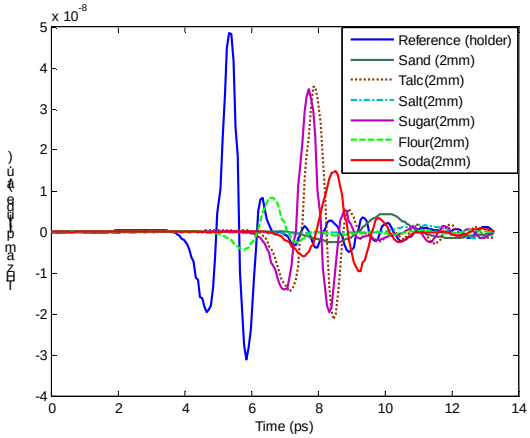


Fig. 2. (a): Illustration of T-rays pulses through 2mm thickness of six different powders and their holder (reference) in the time domain

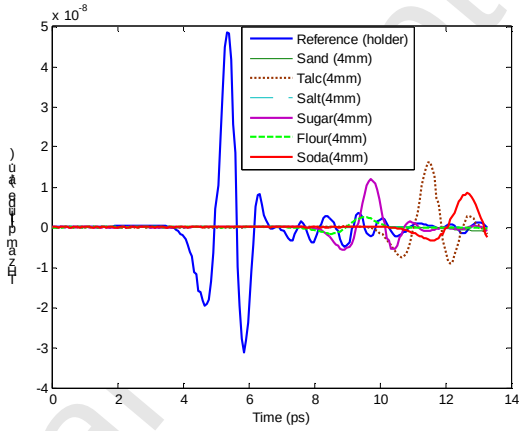


Fig. 2. (b): Illustration of T-rays pulses through 4mm thickness of six different powders and their holder (reference) in the time domain

respectively. It can be seen that the weakest (most attenuated) signals are seen for the powders with sand and salt. According to expectations, as the thickness of the powders were varied the T-rays pulse showed a linear increase in phase (or delay of the time domain pulse) and an exponential decay in amplitude with thickness.

In the proposed methodology, each pixel (a T-ray pulse signal) signal in a powder substance is cross-correlated with the reference signal (holder signal) so that it produces a cross-correlogram

sequence. Each of the six powder substances is composed of a 51 pixels signal irrespective of thickness (e.g. 2mm, 4mm, 3mm, the mixture of 2mm and 4mm, and the mixture of 2mm, 3mm and 4mm thick samples whether they are in pure form or mixture. The reference signal also is composed of 51 pixel signals and each pixel signal contains 401 data points. In the proposed scheme, the reference signal is cross-correlated with the data of a class with the 51 pixel signals using equation (1) and thus for each powdered substance, 51 cross-correlation sequences are obtained where each sequence contains 801 data points. As mentioned in Section 2.1.1, if a reference signal (X) and any other signal (H) of a class consists of S number of samples, the resultant cross-correlation sequence has $2S-1$ samples. Here, S=401.

Hence, each class powder samples corresponds a cross-correlation sequence matrix with dimension 801×51. The proposed 2D cross-correlation approach ensures far superior de-noising than a traditional single pixel by pixel cross-correlation but at the expense of

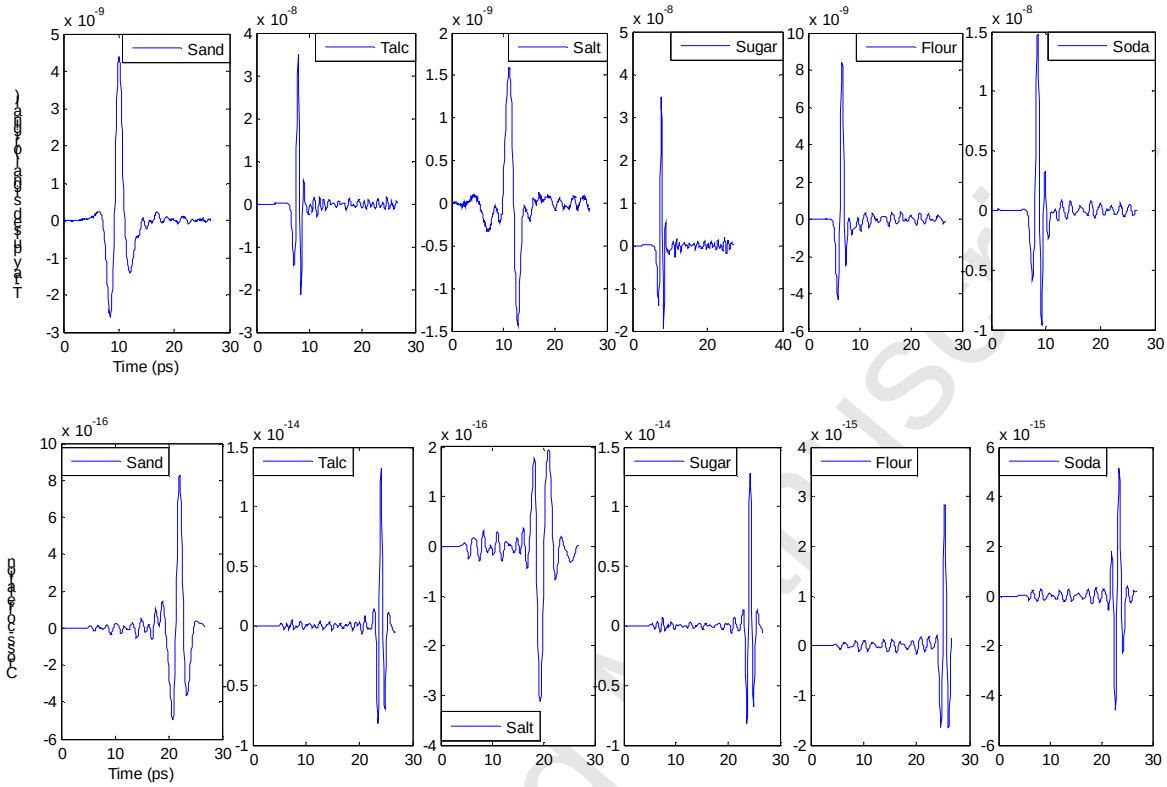


Fig. 3. An example of T-ray signal for 2mm sand, talc, salt, sugar, flour and soda with their corresponding cross-correlation sequence

additional computations. Fig. 3 shows an example of the calculated cross-correlogram patterns. Each cross-correlogram is calculated using equation (1) for each time *lag*. From this figure, one can see that in most of the cases, the shapes of the curves are not exactly the same, this indicates statistical independency.

This pre-processing stage is followed by calculation of the ten statistical parameters (see discussion in Section 2.1.2) from each of the 51 cross-correlation sequences in a class so as to obtain feature matrices with dimension 51×10. Thus, for all six categories of powder data samples, we

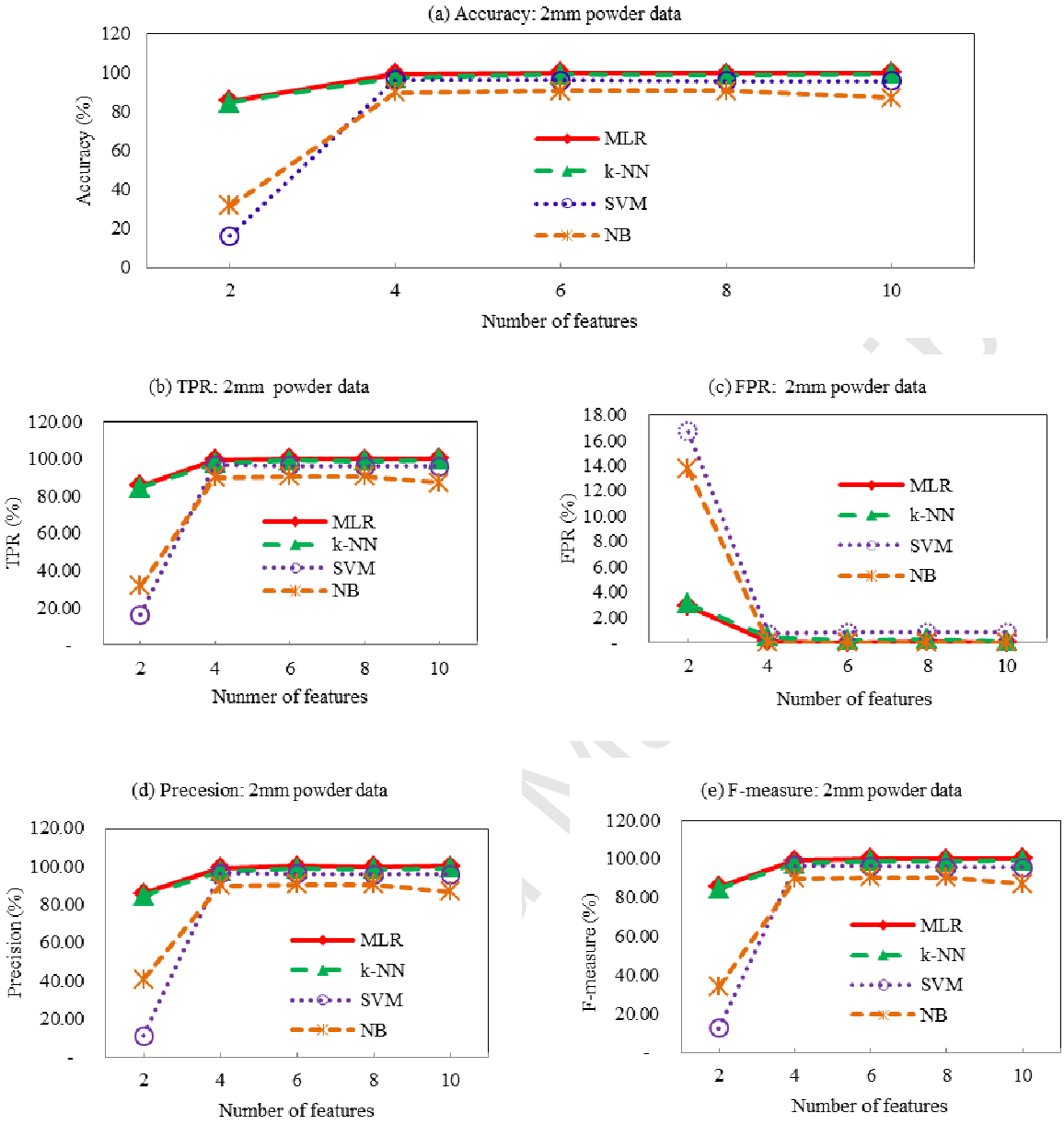


Fig. 4. Classification performances for different number of features on 2mm thickness powder data: (a) accuracy (b) TPR (c) FPR (d) Precision and (e) F-measure.

acquire a total of 306 feature vectors with 10 dimensions. MATLAB functions were employed for calculating mean, standard deviation, skewness, kurtosis, Q_1, Q_3, *IQR*, median, maximum and minimum from each cross-correlation sequence. Using the10-fold cross validation method, the obtained feature vector set is divided into a training set and a testing set. The training set is applied to train the classifier and the testing vectors are used to verify the performances and the effectiveness of

the classifiers. The feature vectors were evaluated through all four classifiers. Classification performances are evaluated in term of accuracy, TPR, FPR, precision and F-measure.

Fig. 4 (a)-(e) shows the variation in performances for the mentioned four classifiers as a function of increased number of input features in the 2mm thickness powder dataset. The number of

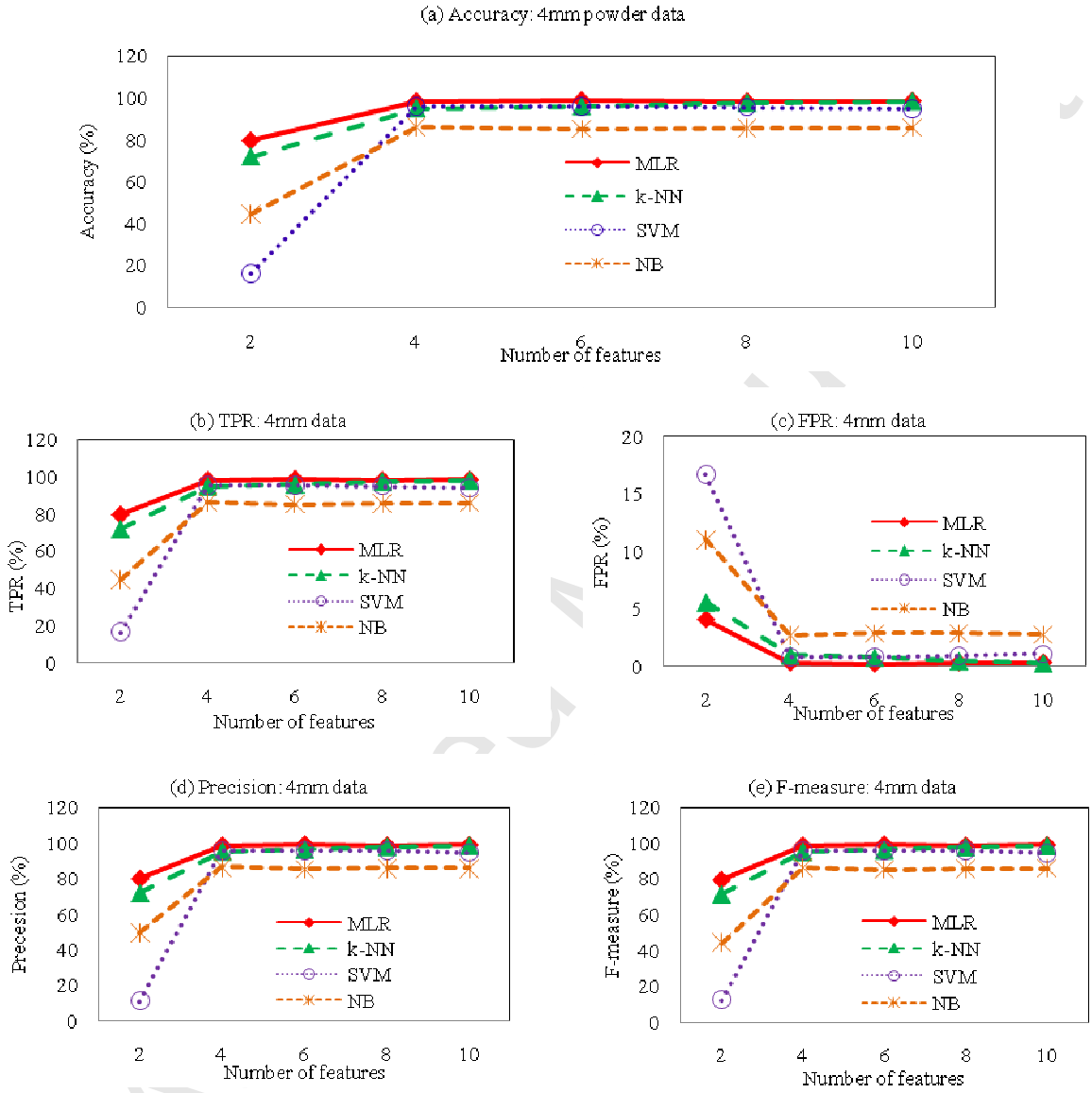


Fig. 5. Classification performances for different number of features on 4mm thickness powder data: (a) accuracy (b) TPR (c) FPR (d) Precision and (f) F-measure.

the input features is varied from 2 to 10. It can be seen that the corresponding accuracy, TPR, precision and F-measure for each four classifiers are increased monotonically and almost linearly with the number of feature vectors and the FPR of each four classifiers are going to decrease with the increase number of feature vectors, this indicates consistency in the proposed analysis. From these figures, it is also observed that in all performance evaluations, the MLR classifier yields a better performance individually, for 2, 4, 6, 8 and 10 features compared to the KNN, SVM and NB

classifiers. As shown in Figs. 4(a)-(e), among the reported four classifiers, the MLR classifier produces the best performances when using 10 features while the NB classifier consistently displays the lowest performances.

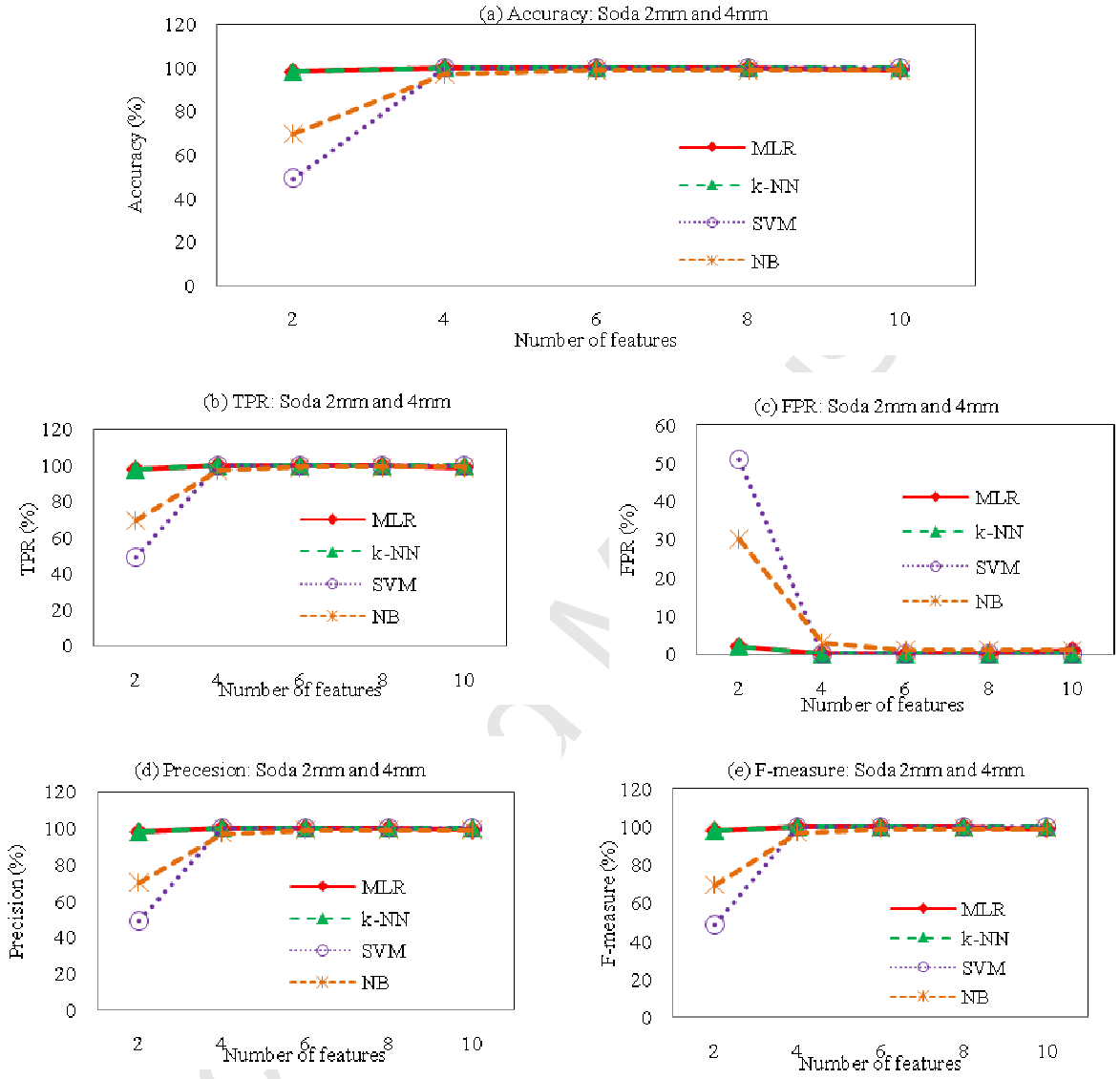


Fig. 6. Classification performances for different number of features on the mixture of 2mm and 4mm soda powder data: (a) accuracy (b) TPR (c) FPR(d) Precision and (e) F-measure.

Figs. 5 (a)-(e) depict the performance of all the classifiers on the basis of the number of features in the 4mm thickness powder sample datasets. Similarly to the results in Fig. 4, the classification performance for each of the four classifiers increases when the number of features is increased. The MLR classifier yields better performance in most of the cases compared to the other three classifiers while the NB classifier performance is the lowest.

Fig. 6 (a)-(e) illustrates the classification accuracy, TPR, FPR, precision and F-measure for all classifiers as a function of number of features for the mixture of 2mm and 4 mm thickness soda powder data. As can be seen, the performance of each of classifiers improves when the number of features considered increases. The highest performances are obtained when assuming 10 features and the lowest for 2 features. In these figures, both MLR and KNN show similar performance, this is superior to that of the other two classifiers on the mixture of 2mm and 4mm thickness soda sample. It can also be seen that the NB classifier is the least successful in the classification task than the other three. This is a very positive overall outcome as it indicates stability consistency and robustness in the results with the 2-D cross correlation feature extraction methodology and the adopted classifier performance evaluation method. These results point to a necessity to use all 10 features for the further evaluation of the proposed classifiers as discussed in the following sections.

4. Results and discussions

Tables 1-3 presents the classification results for all four classifiers in more detail assuming 10 features are used for all powder sample compositions for 2 mm, 4 mm and the mixture of 2mm and 4mm sample thicknesses, respectively. In these three tables, the class-specific performances for each powder substance and also overall performances in terms of accuracy, TPR, FPR, precision and F-measure are reported. In Table 1, it can be observed that the performances (the values of accuracy, TPR, precision and F-measure) for the MLR classifier are most promising, which is 100% across

Table 1: Classification results on 2mm thickness powder data

Classifier	**Performance parameters**	**Classes and their performance (in percentage)**						
		Sand	Talc	Salt	Sugar	Flour	Soda	Overall
MLR	Accuracy	100	100	100	100	100	100	100
	TPR	100	100	100	100	100	100	100
	FPR	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Precision	100	100	100	100	100	100	100
	F-measure	100	100	100	100	100	100	100
KNN	Accuracy	100	98.04	100	98.04	100	100	99.35
	TPR	100	98.00	100	98.00	100	100	99.33
	FPR	0.0	0.40	0.0	0.40	0.0	0.0	0.133
	Precision	100	98.00	100	98.00	100	100	99.33
	F-measure	100	98.00	100	98.00	100	100	99.33
SVM	Accuracy	100	84.31	100	90.20	100	100	95.75
	TPR	100	84.30	100	90.20	100	100	95.75
	FPR	0.0	2.00	0.0	3.10	0.0	0.0	0.85
	Precision	100	89.60	100	85.20	100	100	95.80
	F-measure	100	86.90	100	87.60	100	100	95.75
NB	Accuracy	98.04	64.75	96.08	68.63	96.08	100	87.26
	TPR	98.00	62.70	96.10	68.60	96.10	100	86.92

	FPR	1.20	6.70	0.40	7.10	0.40	0.0	2.63
	Precision	94.30	65.30	98.00	66.00	98.00	100	86.93
	F-measure	96.20	64.00	97.00	67.30	97.00	100	86.92

every category irrespective of powder substance and the FPR is also 0%. Furthermore, the performance parameter values for the KNN classifier is slightly better than those of the SVM and NB classifiers while the SVM classifier performs better than the NB classifier. In addition, the soda powder samples are the easiest to be separated, with classification accuracy of 100% in all cases, whereas the talc and sugar powder samples are the most difficult to classify. The results in Table 1 also clearly shows that the MLR classifier using a10 feature set yields the best performance across all classifiers and the NB classifier shows a consistently inferior performance.

As shown in Table 2, the overall accuracy of the MLR, KNN, SVM and NB classifiers are 98.69%, 98.37%, 95.75% and 87.26%, respectively for the 4mm thickness powder samples on the basis of 10 features being presented at their inputs. The overall TPR for the MLR, KNN, SVM and NB classifiers are 98.7%, 98.37%, 94.45% and 85.95%, respectively and the FPR values are 0.27%, 0.33%, 01.12% and 2.80% respectively. The overall precision and F-measure are 98.7% and 98.68% for the MLR, 98.37% and 98.35% for the KNN, 94.80%, 94.40% for the SVM and 85.87% and

Table 2: Classification results on 4mm thickness powder data

Classifier	**Performance parameters**	**Classes and their performance (in percentage)**						
		Sand	Talc	Salt	Sugar	Flour	Soda	Overall
	Accuracy	100	96.08	100	96.08	100	100	98.69
MLR	TPR	100	96.10	100	96.10	100	100	98.7
	FPR	0.0	0.80	0.0	0.40	0.40	0.0	0.27
	Precision	100	96.10	100	98.00	98.10	100	98.7
	F-measure	100	96.10	100	97.00	99.00	100	98.68
KNN	Accuracy	100	96.08	100	94.12	100	100	98.37
	TPR	100	96.10	100	94.10	100	100	98.37
	FPR	0.0	1.20	0.0	0.80	0.0	0.0	0.33
	Precision	100	94.20	100	96.00	100	100	98.37
	F-measure	100	95.10	100	95.00	100	100	98.35
SVM	Accuracy	100	92.16	100	74.51	100	100	94.45
	TPR	100	92.20	100	74.50	100	100	94.45
	FPR	0.0	5.10	0.0	1.60	0.0	0.0	01.12
	Precision	100	78.30	100	90.50	100	100	94.80
	F-measure	100	84.70	100	81.70	100	100	94.40
NB	Accuracy	100	60.78	100	56.86	100	98.04	85.95
	TPR	100	60.80	100	56.90	100	98.00	85.95
	FPR	0.0	8.60	0.40	7.80	0.0	0.0	2.80
	Precision	100	58.50	98.10	59.20	100	100	85.97
	F-measure	100	59.60	99.00	58.00	100	99.00	85.93

85.93% for the NB. Thus, in most of the cases, the MLR classifier yields the highest performance and the NB lowest one. Moreover, the sand powder samples are easiest to

separate (classification accuracy of 100% across all four classifiers), whereas the talc and sugar powder samples are more challenging to classify.

Table 3 reports the experimental classification outcomes for the mixture of 2mm and 4mm thickness samples for all six powder substances. This classification is performed as a binary process (2 class classification). Here, the 2mm powder substance is considered as one class and the powder substance of 4mm thickness is considered as another class e.g. classification of a 2 mm sand sample

and a 4 mm sand sample. As can be seen from this table, the powder samples of sand, talc, salt, sugar and flour are easiest to be separated by the MLR, KNN and SVM classifiers, (where a classification accuracy of 100% was achieved under all the cases), whereas the soda powder sample proved more difficult to classify. The NB classifier could not classify successfully powder substance mixtures. Also, the soda powder sample was consistently more difficult to classify.

Table 3: Classification results on the mixture of 2mm and 4mm thickness powder data

Classifier	**Performance parameters**	**Classes and their performance (in percentage)**											
		Sand		Talc		Salt		Sugar		Flour		Soda	
		2mm	4mm	2mm	4mm	2mm	4mm	2mm	4mm	2mm	4mm	2mm	4mm
MLR	Accuracy	100	100	100	100	100	100	100	100	100	100	98.04	100
	TPR	100	100	100	100	100	100	100	100	100	100	98.0	100
	FPR	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0
	Precision	100	100	100	100	100	100	100	100	100	100	100	98.10
	F-measure	100	100	100	100	100	100	100	100	100	100	99.0	99.0
KNN	Accuracy	100	100	100	100	100	100	100	100	100	100	100	100
	TPR	100	100	100	100	100	100	100	100	100	100	100	100
	FPR	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Precision	100	100	100	100	100	100	100	100	100	100	100	100
	F-measure	100	100	100	100	100	100	100	100	100	100	100	100
SVM	Accuracy	100	100	100	100	100	100	100	100	100	100	100	100
	TPR	100	100	100	100	100	100	100	100	100	100	100	100
	FPR	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Precision	100	100	100	100	100	100	100	100	100	100	100	100
	F-measure	100	100	100	100	100	100	100	100	100	100	100	100
NB	Accuracy	92.16	100	96.08	98.04	98.04	100	92.16	98.04	96.08	100	100	98.04
	TPR	92.20	100	96.10	98.00	98.00	100	92.20	98.00	96.10	100	100	98.0
	FPR	0.0	7.80	2.0	3.90	0.0	2.0	2.0	7.80	0.0	3.90	2.0	0.0
	Precision	100	92.7	98.00	96.20	100	98.10	97.90	92.60	100	96.20	98.10	100
	F-measure	95.9	96.2	97.00	97.10	99.00	99.00	94.90	95.20	98.00	98.10	99.0	99.0

In order to further demonstrate the effectiveness of the proposed methods, we also apply our methodology on results obtained using 3 mm sample thicknesses and the results are reported in terms of accuracy, TPR, FPR, precision and F-measure. These details are shown in Table 4. It can be seen that the overall accuracy of the MLR, KNN, SVM and NB classifiers with the 10 features set are 96.73%, 97.38%, 95.42% and 89.87%, respectively for the 3mm thickness powder samples. Here, the accuracy of the KNN classifier is a little bit higher than the MLR classifier while it is the lowest for the NB classifier, this result is

reasonably consistent to those obtained by classifying the 2mm and 4mm powder datasets. The other performance criteria show also similar consistency in classification accuracy. Similarly to the case of the 2mm and 4mm thickness sand and soda powder samples, the 3 mm samples are the easiest to be separated, with classification accuracy of 100% in all cases for all four reported classifiers, whereas the talc and sugar powder samples are the most difficult to classify.

Table 4: Classification results on 3mm thickness powder data

Classifier	Performance parameters	Classes and their performance (in percentage)						
		Sand	Talc	Salt	Sugar	Flour	Soda	Overall
MLR	Accuracy	100	88.24	96.08	96.08	100	100	96.73
	TPR	100	88.20	96.10	96.10	100	100	96.70
	FPR	1.20	0.8	0.0	1.60	0.0	0.40	0.70
	Precision	94.40	95.70	100.0	92.50	100.0	98.10	96.80
	F-measure	97.10	91.80	98.0	94.20	100	99.00	96.70
KNN	Accuracy	100.0	90.19	100.0	94.12	100.0	100.0	97.38
	TPR	100.0	90.20	100.0	94.10	100.0	100.0	97.40
	FPR	0.0	1.20	0.0	2.00	0.0	0.0	0.50
	Precision	100.0	93.90	100.0	90.60	100.0	100.0	97.40
	F-measure	100.0	92.00	100.0	92.30	100.0	100.0	97.40
SVM	Accuracy	100.0	78.43	100.0	94.12	100.0	100.0	95.42
	TPR	100.0	78.40	100.0	94.10	100.0	100.0	95.40
	FPR	0.0	1.20	0.0	4.30	0.0	0.0	0.90
	Precision	100.0	93.00	100.0	81.40	100.0	100.0	95.70
	F-measure	100.0	85.10	100.0	87.30	100.0	100.0	95.40
NB	Accuracy	100.0	76.47	100.0	74.51	88.23	100.0	89.87
	TPR	100.0	76.50	100.0	74.50	88.20	100.0	89.90
	FPR	0.0	7.50	0.0	4.70	0.0	0.0	2.00
	Precision	100.0	67.20	100.0	76.00	100.0	100.0	90.50
	F-measure	100.0	71.60	100.0	75.20	93.80	100.0	90.10

Table 5 reports the classification outcomes for the mixture of 2mm, 3mm and 4mm thickness samples for all six powder substances. This classification task is set up as a three class problem. Here, the 2mm thickness powder substance is considered as belonging to the first class, the 3mm thickness powder substance is considered as belonging to the second class and the 4mm thickness powder substance is considered as belonging to the third class. As can be seen from this table, the overall accuracy for the MLR is 99.56% for all the powder samples while this value is 99.35% for KNN, 91.83% for SVM and 91.82% for NB classifier. Similarly to the classification results discussed in the previous sections, in most of the cases, the MLR classifier consistently yields the highest performance whereas the NB classifier the lowest one. As shown in Table 5, the good classification performance and classification consistency of the proposed method in discriminating across samples in a mixture consisting of three thickness (2mm, 3mm and 4mm) powder data sets when from a compositional perspective these samples were originally very hard to discriminate, demonstrate that the 2D cross correlation based feature extraction approach successfully de-noises the datasets while at the same time enables us to resolve useful features in the time domain signals associated with each pixel in the image in a consistent manner. This is significant bearing in mind that classification tasks that were difficult to perform in the past due to the presence of some unquantifiable scattering become now possible. It is also worth noting that although in analytical sciences, cross-correlation techniques have been mainly explored within a de-

noising context, the proposed methodology places these algorithms within a machine learning context. It may also be concluded that the MLR is a powerful and less

Table 5: Classification results on 2mm, 3mm and 4mm thickness powder data

Classifier	Performance parameters	Classification performance (in percentage) among three thickness: 2mm, 3mm and 4mm of each powder																	
		Sand			Talc			Salt			Sugar			Flour			Soda		
		2mm	3mm	4mm	2mm	3mm	4mm	2mm	3mm	4mm	2mm	3mm	4mm	2mm	3mm	4mm	2mm	3mm	4mm
MLR	Accuracy	98.04	98.04	98.04	100	100	100	100	100	100	98.04	100	100	100	100	100	100	100	100
	TPR	98.0	98.0	98.0	100	100	100	100	100	100	98.0	100	100	100	100	100	100	100	100
	FPR	0.0	2.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
	Precision	100	96.2	98.0	100	100	100	100	100	100	100	100	98.10	100	100	100	100	100	100
	F-measure	99.0	97.10	98.0	100	100	100	100	100	100	99.0	100	99.0	100	100	100	100	100	100
KNN	Accuracy	100	96.08	100	100	100	100	100	96.08	98.04	100	100	100	100	100	100	98.04	100	100
	TPR	100	96.10	100	100	100	100	100	96.10	98.0	100	100	100	100	100	100	98.0	100	100
	FPR	0.0	0.0	2.0	0.0	0.0	0.0	0.0	1.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
	Precision	100	100	96.2	100	100	100	100	98.0	96.20	100	100	100	100	100	100	100	98.10	100
	F-measure	100	98.0	98.10	100	100	100	100	97.0	97.10	100	100	100	100	100	100	99.0	99.0	100
SVM	Accuracy	100	98.04	92.16	96.08	98.04	96.08	100	43.14	100	80.39	62.75	100	100	96.08	94.12	100	96.08	100
	TPR	100	98.0	92.2	96.10	98.0	96.10	100	43.10	100	80.40	62.70	100	100	96.10	94.10	100	96.10	100
	FPR	0.0	3.9	1.0	1.0	3.9	0.0	0.0	0.0	28.40	18.6	9.8	0.0	0.0	2.90	2.0	0.0	0.0	2.0
	Precision	100	92.6	97.9	98.0	92.6	100	100	100	63.8	68.30	76.20	100	100	94.20	96.0	100	100	96.20
	F-measure	100	95.2	94.9	97.0	95.20	98.0	100	60.3	77.90	73.90	68.80	100	100	95.10	95.0	100	98.0	98.10
NB	Accuracy	90.19	96.08	94.12	86.27	84.31	78.43	88.24	80.39	90.19	90.19	86.27	94.12	96.08	74.51	96.08	92.16	82.35	74.51
	TPR	90.2	96.10	94.10	86.30	84.30	78.40	88.20	80.40	90.20	90.20	86.30	94.10	96.10	74.50	96.10	92.20	82.40	74.50
	FPR	1.0	2.0	6.9	3.90	14.70	6.90	0.0	9.8	10.8	2.0	2.9	9.8	7.8	2.0	6.9	2.90	16.7	5.9
	Precision	97.90	96.10	87.30	91.70	74.10	85.10	100	80.40	80.70	95.80	93.60	82.80	86.0	95.0	87.50	94.0	71.20	86.40
	F-measure	93.90	96.10	90.60	88.90	78.90	81.60	93.80	80.40	85.20	92.90	89.80	88.10	90.70	83.50	91.60	93.10	76.40	80.0

complex algorithm for THz pulse signals classification. The proposed technique should extend the use of classification algorithms to experiments where samples are not placed in a cuvette, a sample holder or compressed in pellet form in order to perform the spectroscopic investigations, and points towards a new way of performing industrial quality control using THz imaging systems ‘in situ’ when samples are still in powder form where different degree of scattering may also be present in the measurement process across the different spectral bands. The proposed methodology therefore has the potential to significantly extend the applications domain of classifiers for material characterization; this has important applications in high value manufacturing such as the pharmaceutical industry as well as for tissue differentiation and characterization in biomedical imaging.

Fig.7 displays the proposed algorithm execution time for all four classifiers for a 10 feature input across all sample thicknesses. It can be seen that, in every cases the SVM

classifier takes more time than all other reported classifiers and the NB and KNN algorithms are the fastest to execute.

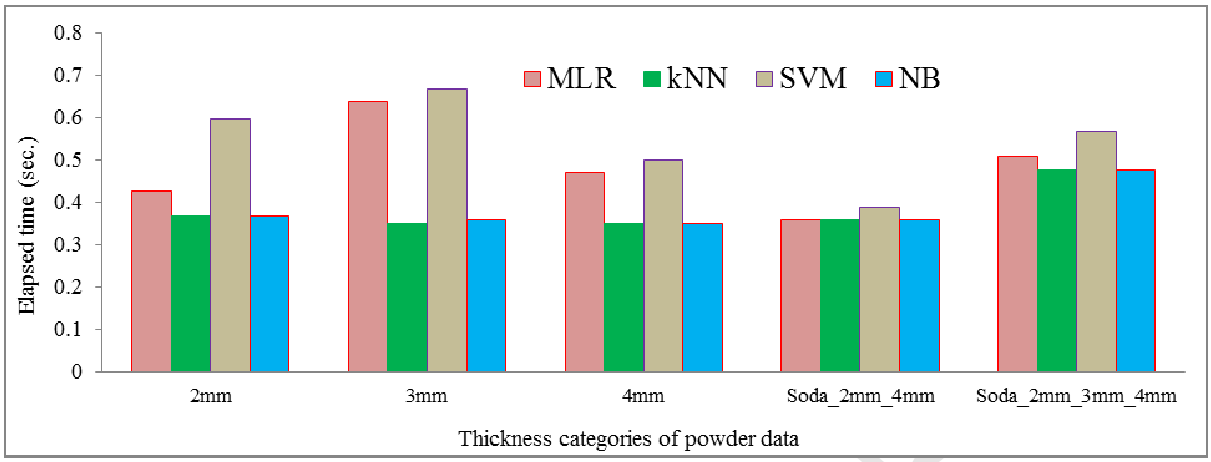


Fig. 7. Elapsed time (in second) for the MLR, KNN, NB and SVM classifiers on 2mm, 3mm and 4mm thickness powder data and the mixture of 2mm & 4mm soda powder sample as well as the mixture of 2mm, 3mm & 4mm soda powder sample dataset.

The shape of the MAE for each of the four reported classifiers is illustrated in Fig.8. The lower MAE score indicates a higher performance in the proposed approach. We can see that irrespective of thickness the score of MAE is significantly lower for the MLR classifier compared to the other three classifiers. On the other hand, the NB classification method consistently yields a very high MAE score. Particularly, in the cases of mixture of 2mm, 3mm and 4mm powder samples, both

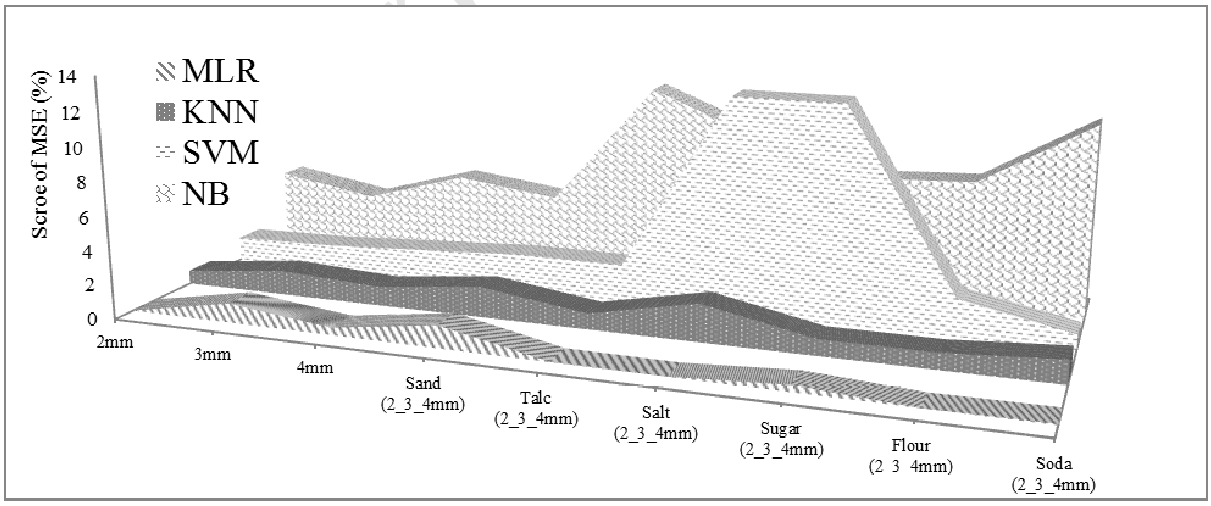
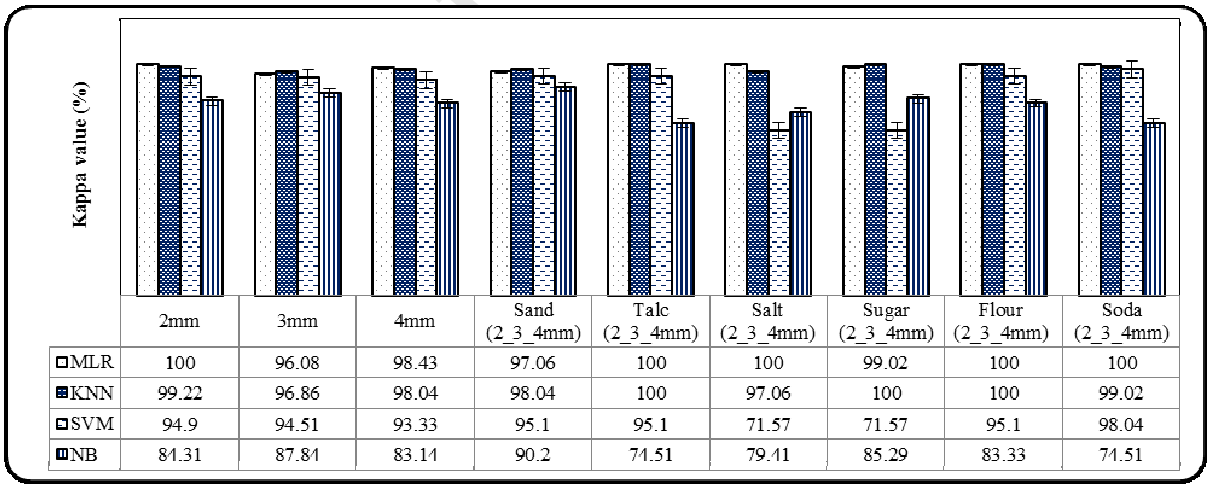


Fig. 8. 3-D stacked area graph showing MAE score for the MLR, KNN, SVM and NB classifiers for different thickness powder samples.

NB and SVM generate very high MAE scores while the values are very low for the MLR classifier. Once again the MLR classifier seems to be the best choice to classify the THz pulses signal datasets associated with different powder compositions.

Fig. 9 displays kappa statistics for all classifiers assuming a 10 feature input. The aim of the kappa statistics test is to evaluate the consistency of the classifiers. Consistency is considered mild if kappa values are less than 0.2 (20%), fair if it lies between 0.21-0.40 (21-40%), moderate if it lies between 0.41-0.60 (41-60%), good if it is between 0.61-0.80 (61-80%), and excellent if it is greater than 0.81 (81%). As shown in Fig.9, the highest kappa values are obtained by the MLR on both 2mm thickness sample datasets (100%), as well as 4mm (98.43%) datasets. In addition, highest kappa values are obtained for the mixture of 2mm, 3mm and 4mm samples of talc (100%), salt (100%), flour (100%) and soda (100%). The KNN algorithm also demonstrated very good performance (second best overall) as can be seen in the case of the 3mm thickness sample datasets (96.86%), and the mixtures of 2mm, 3mm and 4mm sand (98.04%), talc (100%), sugar (100%) and flour (100%). The kappa values of the other two classifiers (SVM and NB) are systematically lower compared to those achieved by the MLR and KNN irrespective of sample type, furthermore the values are consistently lowest for the NB classifier. In this figure, the error bars indicate the associated kappa value standard error. In most of the cases, the highest kappa values are obtained using the MLR algorithm.



	2mm	3mm	4mm	Sand (2_3_4mm)	Talc (2_3_4mm)	Salt (2_3_4mm)	Sugar (2_3_4mm)	Flour (2_3_4mm)	Soda (2_3_4mm)
MLR	100	96.08	98.43	97.06	100	100	99.02	100	100
KNN	99.22	96.86	98.04	98.04	100	97.06	100	100	99.02
SVM	94.9	94.51	93.33	95.1	95.1	71.57	71.57	95.1	98.04
NB	84.31	87.84	83.14	90.2	74.51	79.41	85.29	83.33	74.51

Fig. 9. Kappa statistics values for the MLR, KNN, SVM and NB classifiers for datasets associated with different powder thickness samples.

In order to compare our research outcomes with existing ones in the literature the only reference that can be found is discussed in [24]. In their work, however, the focus on the study was

placed on the derivation of hybrid AR and ARMA models with further wavelet compression for very parsimonious feature extraction aiming to improve on the generalization ability of the classifier. In that study, wavelet-based de-noising with soft threshold shrinkage was applied to the measured T-ray signals prior to modeling. It is also worth noting that a simple Mahalanobis distance classifier was used at that time for the classification of the powder samples and the emphasis was placed on feature extraction as opposed to address state-of-the-art machine learning approaches. An overall 98% classification accuracy for all thickness powders was achieved with that approach whereas the proposed method based on 2D-cross-correlation and an MLR classifier yielded a classification accuracy of 99.56%.

In this study, in most of the circumstances, the MLR algorithm produced better results compared to other reported three classifiers and the total performance of the KNN classifier was alike to the MLR classifier. As mentioned before, T-rays pulse signals contains multi-correlation in different powder substance data because screening items are often highly correlated in terms of particle shape and dimension, the effect can also manifest itself at specific measurement angles due to possible scattering. The one of the main advantages of the MLR is to properly handle multicollinearity within a large number of covariates that cause unstable in the parameter estimation and larger variance in the associated distributions used as inputs to the classifiers. These effects collectively have an overall effect of systematically degrading classification accuracy. One drawback of the MLR may be the computational demand needed when images are composed of a very large number of pixels. On the other hand, the key advantage of the KNN classifier is, it does not require a priori assumptions regarding the distributions from which the training examples are drawn. It makes this method to be simple in implementation with less computation time. But the core limitation of this method is, it's classifying accuracy decreases in the presence of high dimensional feature data. Hence, it seems that from the current study, the MLR promises to offer the better performance for detection of powder substance using T-rays pulse signal datasets reducing overfitting error.

5. Conclusions

This paper presented the first systematic evaluation of machine learning algorithms tailored specifically to the classification of THz datasets obtained using a THz transient spectrometer. A further aim was to establish alternative new criteria that would capture some of the features present in the time-domain signals so that unquantifiable scattering effects that would otherwise degrade the discriminating ability of the classifiers would be minimized. A 2-D

cross correlation technique was adopted for feature extraction prior to sample classification. The dimensions of the calculated cross-correlation sequences were also further compressed extracting additional statistical features. Several powder substances of various thickness and composition were successfully classified using the proposed algorithms. Systematic evaluation of the performance of the four classifiers considered using multiple datasets of powdered samples and a comprehensive cross-validation methodology showed that, in most of the cases, the MLR classifier with ridge estimator outperformed the KNN, SVM and NB classifiers. It is worth noting however, that the overall performance of the KNN classifier was very similar to that of the MLR classifier. Thus the study concluded that the 2-D cross-correlation based MLR or KNN algorithm in conjunction with the proposed 2-D cross-correlation technique can lead to a systematic enhancement in THz transient dataset classification success rate. The proposed methodology paves the way for establishing new robust and consistent approaches for the analysis and automated classification of THz transient biomedical imaging datasets which are currently difficult to classify because of the large signal attenuation of tissue associated with the quenching from the tissue's water content. Algorithmic expert systems are currently considered to be the Achilles' heel in THz signal analysis. Thus the work addresses a fundamental problem which so far has consistently delayed the further proliferation and commercialization of THz transient spectrometers. Future investigations will assess classifier performance when the THz transient data vectors are systematically over-sampled or under-sampled (always above the Nyquist's criterion), so as to assess opportunities for optimizing signal to noise ratio for a given data acquisition time frame, this is a topic of significant interest to clinicians considering the adoption of this imaging modality for routine patient screening. Beyond the THz community, the proposed methodology may also be used for the systematic assessment of different classifiers and automated expert systems as applied to other datasets across the entire electromagnetic spectrum e.g. X-ray, UV, visible, infrared or microwave spectrometry, electron spin resonance spectrometry, nuclear magnetic resonance imaging, positron emission tomography etc. Thus the work is generic and of relevance across all physical sciences.

References

[1] B. Ferguson and X.-C. Zhang, "Materials for terahertz science and technology," *Nature Materials* **1**(1), pp. 26–33, 2002.

[2] M. C. Nuss, Chemistry is right for T-rays, *IEEE Circuits and Devices* 12(2), pp. 25-30, 1996

[3] W. Withayachumnankul et al., T-Ray Sensing and Imaging, *Proceedings of the IEEE*, 2007; 95 (8):1528-1558.

[4] X. Yin, B. Ng, D. Abbott, Terahertz Imaging for Biomedical Applications: *Pattern Recognition and Tomographic Reconstruction*, Springer-Verlag, New York, 2012.

[5] J. A. Zeitler, Philip F. Taday, David A. Newnham, M. Pepper, K. C. Gordon and T. Rades 'Terahertz pulsed spectroscopy and imaging in the pharmaceutical setting – a review', *Pharmacy and Pharmacology*, 59, 209-223, (2007)

[6] C. J Strachan, P. F Taday, D. A Newnham, K. C Gordon, J A. Zeitler, M. Pepper, T. Rades, 'Using terahertz pulsed spectroscopy to quantify pharmaceutical polymorphism and crystallinity', *Journal of Pharmaceutical Sciences* 94 (4), 837-846, (2005)

[7] J Axel Zeitler, Y. Shen, C. Baker, P. F Taday, M. Pepper, T. Rades, 'Analysis of coating structures and interfaces in solid oral dosage forms by three dimensional terahertz pulsed imaging', *Journal of Pharmaceutical Sciences* 96 (2), 330-340, (2007).

[8] L. Ho, R. Müller, K. C. Gordon, P. Kleinebudde, M. Pepper, T. Rades, Y. Shen, P. F Taday, J A. Zeitler, 'Monitoring the film coating unit operation and predicting drug dissolution using terahertz pulsed imaging', *Journal of Pharmaceutical Sciences* 98 (12), 4866-4876, (2009).

[9] D. Brock, J A. Zeitler, A. Funke, K. Knop, P. Kleinebudde, 'Evaluation of critical process parameters for intra-tablet coating uniformity using terahertz pulsed imaging', *European Journal of Pharmaceutics and Biopharmaceutics*, 85(3), 1122-1129 (2013).

[10] P. Bawuah, A. P. Mendia, P. Silfsten, P. Pääkkönen, T. Ervasti, J. Ketolainen, J A. Zeitler, K.-E. Peiponen, *International journal of pharmaceutics*, 465 (1), 70-76, (2014).

[11] J A Zeitler, D A Newnham, P F Taday, T L Threlfall, R W Lancaster, R W Berg, C J Strachan, M Pepper, K C Gordon, T Rades, 'Characterization of temperature-induced phase transitions in five polymorphic forms of sulfathiazole by terahertz pulsed spectroscopy and differential scanning calorimetry' *Journal of Pharmaceutical Sciences* 95 (11), 2486-2498, (2006)

[12] E. P.J. Parrott, J. A. Zeitler, T. Friščić, M. Pepper, W. Jones, G. M. Day, L. F Gladden, 'Testing the sensitivity of terahertz spectroscopy to changes in molecular and supramolecular structure: a study of structurally similar cocrystals', *Crystal Growth and Design*, 9 (3), 1452-1460, (2009).

[13] R. Li, J A. Zeitler, D. Tomerini, E. P.J. Parrott, L. F Gladden, G. M Day, 'A study into the effect of subtle structural details and disorder on the terahertz spectrum of crystalline benzoic acid', *Physical Chemistry Chemical Physics*, 12 (20), 5329-5340, (2010).

[14] J A. Zeitler, P. F Taday, M. Pepper, T. Rades, 'Relaxation and crystallization of amorphous carbamazepine studied by terahertz pulsed spectroscopy', *Journal of Pharmaceutical Sciences* 96 (10), 2703-2709, (2007).

[15] J. Sibik, M. J Sargent, M. Franklin, J A. Zeitler, 'Crystallization and phase changes in paracetamol from the amorphous solid to the liquid phase', *Molecular pharmaceutics*, 11 (4), 1326-1334, (2014)

[16] M. Haaser, K. Naelapää, K. C. Gordon, M. Pepper, J. Rantanen, C. J Strachan, P. F. Taday, J. A. Zeitler, T. Rades, 'Evaluating the effect of coating equipment on tablet film quality using terahertz pulsed imaging', *European Journal of Pharmaceutics and Biopharmaceutics*, 85 (3), 1095-1102, (2013),

[17] J. Sibik, K. Lobmann, T. Rades, J.A. Zeitler, 'Predicting crystallization of amorphous drugs with terahertz spectroscopy', *Molecular pharmaceutics*, 12(8), 3062-3068, (2015).

[18] J. A. Zeitler, L. F Gladden, 'In-vitro tomography and non-destructive imaging at depth of pharmaceutical solid dosage forms', *European Journal of Pharmaceutics and Biopharmaceutics*, 71, (1), 2-22, (2009)

[19] P. Bawuah, P. Silfsten, T. Ervasti, J. Ketolainen, J. A. Zeitler, K.-E. Peiponen, 'Non-contact weight measurement of flat-faced pharmaceutical tablets using terahertz transmission pulse delay measurements', *International journal of pharmaceutics*, 476(1), 16-22, (2014).

[20] D. Brock, J. A. Zeitler, A. Funke, K. Knop, P. Kleinebudde, 'Evaluation of critical process parameters for inter-tablet coating uniformity of active-coated GITS using Terahertz Pulsed Imaging', *European Journal of Pharmaceutics and Biopharmaceutics*, 88(2), 434-442, (2014).

[21] N. Y Tan, J. A. Zeitler 'Probing Phase Transitions in Simvastatin with Terahertz Time-Domain Spectroscopy', 12(3), 810-815, 2015, D. Brock, J.A. Zeitler, A. Funke, K. Knop, P. Kleinebudde, 'A comparison of quality control methods for active coating processes', *International journal of pharmaceutics*, 439, (1) 289-295, (2012).

[22] J. A. Zeitler, Y.-C. Shen, 'Industrial applications of terahertz imaging', *Terahertz Spectroscopy and Imaging* 451-489, Springer Berlin Heidelberg (2012).

[23] C. Strachan, P. Taday, D. Newnham, K. Gordon, J. Zeitler, M. Pepper, T. Rades, Using terahertz pulsed spectroscopy to quantify pharmaceutical polymorphism and crystallinity, *Opt. Express* 94 (4) (2005) 837–846.

[24] X. Yin, B. Ng, and D. Abbott, Application of auto regressive models of wavelet sub-bands for classifying terahertz pulse measurements, *Journal of Biological Systems*, 15 (4) (2007) 551–571.

[25] J. Federici, B. Schulkin, F. Huang, D. Gary, R. Barat, F. Oliveira, D. Zimdars, THz imaging and sensing for security applications-explosives, weapons and drugs, *Semicond. Sci. Technol*. 20 (7) (2005) S266–S280.

[26] K. Kawase, Y. Ogawa, Y. Watanabe, H. Inoue, Non-destructive terahertz imaging of illicit drugs using spectral fingerprints, *Opt. Express* 11 (20) (2003) 2549–2554.

[27] T. Ding, R. Li, J. A. Zeitler, T. L. Huber, L. F Gladden, A. P.J. Middelberg, R. J. Falconer, 'Terahertz and far infrared spectroscopy of alanine-rich peptides having variable ellipticity', *Optics Express*, 18 (26), 27431-27444, 2010

[28] S. Hadjiloucas and J.W. Bowen 'Precision of Quasi-optical Null-Balanced Bridge Techniques for Transmission and Reflection Coefficient Measurements,' *Review of Scientific Instruments*, 70, 213-219 (1999)

[29] JW Bowen, GC Walker, S Hadjiloucas, E Berry, 'The consequences of diffractively spreading beams in ultrafast THz spectroscopy', Joint 29th International Conference on Infrared and Millimeter Waves and 12th International Conference on Terahertz Electronics, Karlsruhe, Germany, *IEEE Catalog Number* 04EX857, ISBN: 0-7803-8490-3 pp. 551-552, (2004)

[30] G. M. Hieftje, R. I. Bystroff and Robert Lim, "Application of correlation analysis for signal-to-noise enhancement in flame spectrometry: use of correlation in determination of rhodium by atomic fluorescence," *Analytical Chemistry*, vol. 45, no. 2, pp. 253-258, 1973.

[31] S. Dutta, A. Chatterjee and S. Munshi, Correlation techniques and least square support vector machine

combine for frequency domain based ECG beat classification, *Medical Engineering and Physic*s 32 (2010) 1161-1169

[32] Siuly, Li, Y., and Wen, P. (2014) 'Modified CC-LR algorithm with three diverse feature sets for motor imagery tasks classification in EEG based brain computer interface', *Computer Methods and programs in Biomedicine*, vol. 113, no. 3, pp. 767-780.

[33] Walker, G. C., Bowen, J. W., Labaune, J., Jackson, J.-B., Hadjiloucas, S., Roberts, J., Mourou, G. and Menu, M. (2012) *Terahertz deconvolution.* Optics Express, 20 (25), pp. 27230-27241.

[34] S. Hadjiloucas, R. K. H. Galvão, V. M. Becerra, J. W. Bowen, R. Martini, M. Brucherseifer, H. P. M. Pellemans, P. Haring Bolívar, H. Kurz, J. M. Chamberlain, 'Comparison of state space and ARX models of a waveguide's THz transient response after optimal wavelet filtering,' *IEEE Transactions on Microwave Theory and Techniques MTT*, 52, (10), pp. 2409-2419 (2004).

[35] R.K.H. Galvão, S. Hadjiloucas, V.M. Becerra and J.W. Bowen, 'Subspace system identification framework for the analysis of multimoded propagation of THz-transient signals,' *Measurement Science and Technology*, 16, pp. 1037-1053, (2005).

[36] R.K.H. Galvão S. Hadjiloucas, J.W. Bowen and C.J. Coelho, 'Optimal discrimination and classification of THz spectra in the wavelet domain,' *Optics Express*, 11, 1462-1473 (2003)

1.1.1 [37] B. Ferguson, S. Wang, H. Zhong, D. Abbottand X.-C.Zhang, Powder retection with T-ray imaging, *Proceedings of SPIE Volume 5070, Terahertz for Military and Security Applications*, 2003.

[38] X.-X. Yin, S. Hadjiloucas, Y. Zhang, Complex extreme learning machine applications in terahertz pulsed signals feature sets, *Computer Methods and Programs in Biomedicine*, 1 1 7 (2 0 1 4) 387–403.

[39] X. Yin, B. Ng, B. Fischer, B. Ferguson, D. Abbott, Support vector machine applications in terahertz pulsed signals feature sets, IEEE Sens. J. 7 (12) (2007) 1597–1608.

[40] S. Hadjiloucas, R.K.H. Galvão, J.W. Bowen, R. Martini, M. Brucherseifer, H.P.M. Pellemans, P. Haring Bolivar, H. Kurz, J. Digby, G.M. Parkhurst, J.M. Chamberlain. 'Measurement of propagation constant in waveguides using wideband coherent THz spectroscopy,' *Journal of the Optical Society of America* B 20, 391-401, (2003).

[41] X. X. Yin, K.M. Kong, J.W. Lim, B.W.-H. Ng, B. Ferguson, S.P. Mickan, D. Abbott, Enhanced t-ray signal classification using wavelet preprocessing, Med. Biol. Eng. Comput. 45 (6) (2007)611–616.

[42] Hanbay, D. (2009) 'An expert system based on least square support vector machines for diagnosis of the valvular heart disease', *Expert System with Applications*, Vol. 36, pp.4232-4238.

[43] S. Dutta, A. Chatterjee and S. Munshi, Correlation techniques and least square support vector machine combine for frequency domain based ECG beat classification, *Medical Engineering and Physic*s 32 (2010) 1161-1169

[44] Siuly, Li, Y., and Wen, P. (2014) 'Modified CC-LR algorithm with three diverse feature sets for motor imagery tasks classification in EEG based brain computer interface', *Computer Methods and programs in Biomedicine*, vol. 113, no. 3, pp. 767-780.

[45] R.K.H. Galvão S. Hadjiloucas, J.W. Bowen and C.J. Coelho, 'Optimal discrimination and classification of THz spectra in the wavelet domain,' *Optics Express*, 11, 1462-1473 (2003)

[46] M. N. Islam, *An introduction to statistics and probability*, 3rd ed., Mullick& brothers, Dhaka New Market, Dhaka-1205, pp. 160-161.

[47] R. D. De Veaux, P.F. Velleman and D.E. Bock, *Intro Stats*, 3rd ed., Pearson Addison Wesley, Boston, 2008.

[48] Siuly, Li, Y. (2012) 'Improving the separability of motor imagery EEG signals using a cross correlation-based least square support vector machine for brain computer interface', *IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 20, no. 4, 526-538.*

[49] Siuly, E. Kabir, H. Wang, and Y. Zhang, Exploring Sampling in the Detection of Multicategory EEG Signals, *Computational and Mathematical Methods in Medicine*, Volume 2015, Article ID 576437, 12 pages, http://dx.doi.org/10.1155/2015/576437

[50] Chaovalitwongse, W.A., Fan Y.J., and Sachdeo, R.C. (2007) 'On the Time Series *K*-Nearest Neighbor Classification of Abnormal Brain Activity', *IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans*, vol. 37, no. 6, PP. 1005-1016.

[51] Efron, B. (1983) 'Estimating the error rate of a prediction rule: Improvement on cross-validation,' *J. Amer. Stat. Assoc.*, vol. 78, no. 382, pp. 316–331.

[52] Le Cessie, S., Van Houwelingen, J.C. (1992). Ridge Estimators in Logistic Regression, Applied Statistics. 41(1):191-201.

[53] Zahid F.M., Tutz, G. (2013) Ridge estimation for multinomial logit models with symmetric side constraints, *Comput Stat*, vol. 28, pp. 1017–1034

[54] R. O. Duda, P. E. Hart, & D. G. Strok, *Pattern classification (2nd ed.)*, John , Wiley & Sons, 2001.

[55] Song Y., Huang J., Zhou D., Zha H. and Giles C. L., (2007) I*K-NN*: Informative K-Nearest Neighbor Pattern Classification, *PKDD 2007, LNAI 4702*, pp. 248–264.

[56] Han, J., Kamper. M., Pei, J., *Data mining: Concepts and techniques*, *Morgan Kaufmann*, 2005.

[57][35] Ripley B, *Pattern recognition and neural networks*, Cambridge: Cambridge university press, 1996.

[58] Vapnik, V., The nature of statistical learning theory, Springer-Verlag New York Inc, 2000.

[59] Begg, R. K.,Palaniswami, M.,and Owen, B. (2005) Support Vector Machines for Automated Gait Classification, *IEEE Transactions on Biomedical Engineering*, vol. 52, no. 5.

[60]Mitchel, T., *Machine Learning*, McGraw-Hill Science, 1997.

[61] Wiggins, M., Saad, A., Litt, B. and Vachtsevanos, G. (2011) 'Evolving a Bayesian Classifier for ECG-based Age classification in Medical Applications', *Appl Soft Comput*, Vol. 8, no. 1, 599-608

[62] Bhattacharyya, S. et al. (2011) 'Performance Analysis of Left/Right Hand Movement Classification from EEG Signal by Intelligent Algorithms', *Computational Intelligence, Cognitive Algorithms, Mind, and Brain (CCMB) IEEE Symposium*, 2011.

[63] Frank, E., Hall, M., Holmes, G., Kirkby, R., Pfahringer, B., Witten, I., Trigg, L., (2010) 'Weka-a machine learning workbench for data mining', *Data Mining and Knowledge Discovery Handbook*, pp. 1269–1277.

[64] Chang, C.C. and Lin, C.J., (2011) 'LIBSVM: a library for support vector machines', *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, Article 27, Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

[65] Siuly, and Li, Y. (2014) 'A novel statistical framework for multiclass EEG signal classification', *Engineering Applications of Artificial Intelligence*, vol. 34, pp. 154–167

[66] Faul, S. and Marnane, W., (2012) 'Dynamic, location-based channel selection for power consumption reduction in EEG analysis', Computer Methods and Programs in Biomedicine, vol. 108, pp. 1206-1215.

[67] Siuly, and Li, Y. (2014) 'Discriminating the brain activities for brain computer interface applications through the optimal allocation based approach', *Neural Computing and Applications*, DOI :10.1007/s00521-014-1753-3.

[68] Patnaik, L.M., Manyamb, O. K. (2008) 'Epileptic EEG detection using neural networks and post-classification', Computer Methods and Programs in Biomedicine, vol. 9 1,pp. 100–109.

[69] Siuly, Y. Li,Designing a robust feature extraction method based on optimum allocation and principal component analysis for epileptic EEG signal classification, *Computer Methods and Programs in Biomedicine*, 1 1 9 (2 0 1 5) 29–42

[70] Fraiwan, L., Lweesy, K.; Khasawneh, N., Fraiwan, M., Wenz, H., Dickhaus, H., (2010) 'Classification of Sleep Stages Using Multi-wavelet Time Frequency Entropy and LDA', *Methods Inf Med.*, vol. 49, pp. 230–237.